Appendix E
Statistical Analysis

E.1. Overview.

This appendix describes the statistical procedures that are generally used by the USACE for certification of PE samples. It will not cover the details of experimental design, the derivation of equations, or the justification for use of particular equations. These topics are beyond the scope of this Appendix and are very well covered by a number of standard textbooks.

Statistical procedures along with examples discussed in this Appendix include:

- Distribution test
- Outlier test
- Homogeneity test
- Stability test
- Reproducibility
- Reference value
- Prediction interval

E.2. General Guidelines.

- Use a statistician to assist with all aspects of experimentation.
- Provide the statistician with background information on the proposed experiment so the experimenter and the statistician may determine the best approach to suit the experimental objectives and constraints.
- Consider using statistical tools such as commercially available spreadsheets (e.g., Excel[1]) and statistical software packages (e.g., DataQUEST[2], Minitab[3], SAS[4], etc.) for data handling and analysis.

E.3. Distribution Test.

E.3.1. Purpose. Many statistical models and tests are only appropriate for data that follow a normal distribution or can be transformed into a normal distribution. This is especially true for chemical data. Testing for normality is crucial in selection of appropriate statistical methods.

---

[1] Excel is a registered trademark of Microsoft Corporation, One Microsoft Way, Redmond, WA 98052.
[2] DataQUEST is a free software developed by USEPA, 401 M Street, SW, Washington, DC 20460.
[3] Minitab is a registered trademark of Minitab Inc., 3081 Enterprise Drive, State College, PA 16801.
[4] SAS is a registered trademark of the SAS Institute Inc., Campus Drive, Cary, NC 27513.

E.3.2. Test Types. The test of data distribution can be performed either qualitatively using a normal probability plot, histogram, or stem-and-leaf plot, or quantitatively using a statistical analysis to confirm or reject the assumptions that accompany a statistical test. For normally distributed data, a normal distribution plot approximately follows a straight line. If data are not normally distributed, there are large deviations from a straight line in the tails or middle of a normal distribution plot. Both histogram and stem-and-leaf plots of a normal distribution show bell-shaped curves. Using a plot to decide if the data are normally distributed involves making a subjective decision which is easy to make for extremely non-normal data. When there is no straightforward decision; however, formal quantitative procedures such as W test or Filliben's statistic are usually necessary to test the assumption of normality. W test and Filliben's statistic compute the correlations between the quantiles of the standard normal distribution and the ordered values of sample data. If the data follow a normal distribution curve, the test statistic will be relatively high. However, both tests are difficult to compute manually due to a large number of summations and multiplications. Use USEPA's DataQUEST software to perform both tests. The rest of this section explains the procedure for a normal distribution plot.

E.3.3. Normal Probability Plot.

E.3.3.1. Procedure. Follow these steps to perform a normal probability plot:

1. Order the data ($X_1, X_2, \ldots, X_n$) from the lowest to the highest.
2. Compute the absolute frequency (AF_i), i.e., the number of times each value occurs, for each data value; and the cumulative frequency (CF_i) and Y_i with the following formulas:

$$CF_i = \sum_{i=1}^{n} AF_i$$

$$Y_i = 100 \times \frac{CF_i}{(n + 1)}$$

3. Plot (Y_i, X_i) pairs on a normal probability paper. If the plot of these pairs approximately forms a straight line, the data are probably normally distributed. Otherwise, the data may not be normally distributed.

E.3.3.2. Example. Consider the following eleven data points: 83.0, 86.9, 86.2, 89.7, 80.7, 83.0, 84.1, 87.8, 88.5, 86.2, and 84.5 mg/kg. Rank the data in order and compute the frequencies as shown in Table E-1.

Table E-1. Example of Normal Probability Plot

i	Individual X_i	Absolute Frequency, AF_i	Cumulative Frequency, CF_i	Y_i
1	80.7	1	1	8.33
2	83.0	2	3	25.00
3	84.1	1	4	33.33
4	84.5	1	5	41.67
5	86.2	2	7	58.33
6	86.9	1	8	66.67
7	87.8	1	9	75.00
8	88.5	1	10	83.33
9	89.7	1	11	91.67

A plot of the (Y_i, X_i) pairs using normal probability paper is shown in Figure E.1. Because these pairs apparently form a straight line, the data are probably normally distributed.

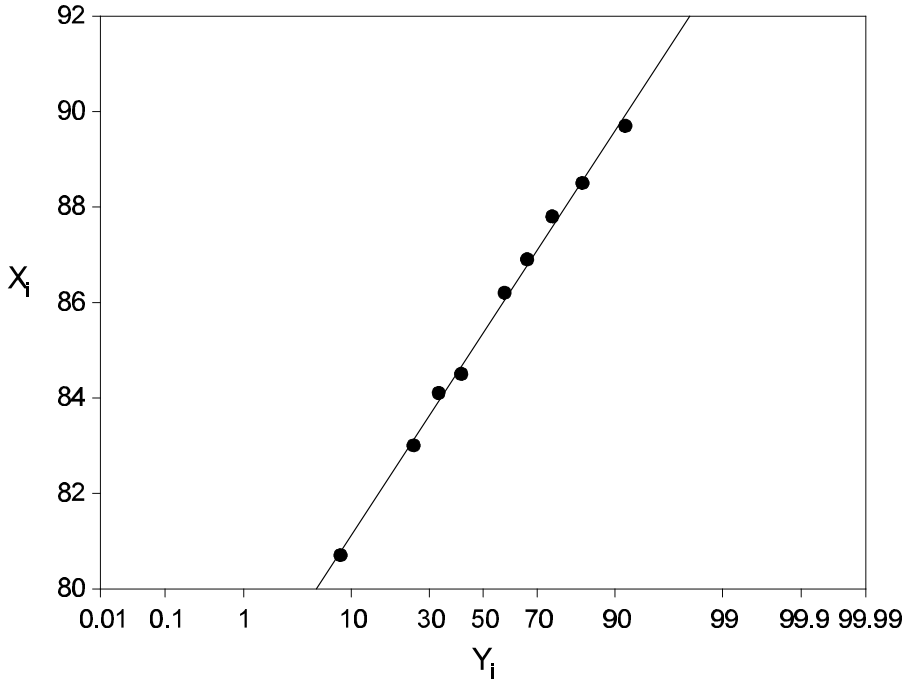

Figure E-1. Normal Probability Plot

E.4. Outlier Test.

E.4.1. Purpose. Outliers are measurements that are extremely large or small compared with the rest of the sample data and are suspected of misrepresenting the population from which they were collected. Statistical outlier tests provide probabilistic evidence that the extreme values do not fit with the distribution of the remainder of the data and are therefore statistical outliers. These tests should only be used to identify data points that require further investigations. The tests alone cannot determine if a statistical outlier should be discarded or corrected. This decision should be based on scientific and judgmental grounds, in addition to the results of statistical outlier tests.

E.4.2. Test Types. USEPA's DataQUEST software package provides several popular tests including Grubbs' tests, Dixon's test, Rosner's test, and Walsh's test. Select a test based on sample size, data distribution, the number of outliers, etc. Follow these guidelines also:

- Use Grubbs' test when data are normally distributed and the sample size is not greater than 50.
- Use Rosner's test if sample size is equal to or greater than 50.
- Use a nonparametric test, such as Walsh's test, if the data are not normally distributed or cannot be transformed.

An example calculation of the commonly used Grubbs' test is presented below.

E.4.3. Grubbs' Test.

E.4.3.1. Procedure. Three similar tests to predict outliers in normally distributed data were developed by Frank Grubbs. The specific Grubbs' test usually used by the USACE PE Program considers the smallest and/or largest value(s) of the data set as the suspected outlier(s) and is discussed here.

1. Perform a normality test of underlying data distribution without the suspected outliers prior to performing a Grubbs' test.
2. If data pass a normality test, rank data from the smallest to the largest to detect any suspected outliers.
3. Compute the sample mean and standard deviation (*SD*) according to the following formulas:

$$mean = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$SD = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}}$$

$$= \sqrt{\frac{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2}{n - 1}}$$

If the suspected outlier is the smallest value of the data set, the test statistic of the Grubbs' test is:

$$G = \frac{\bar{X} - X_1}{SD}$$

If the suspected outlier is the largest value of the data set, the test statistic is:

$$G = \frac{X_n - \bar{X}}{SD}$$

If G exceeds the critical value in Table F-1 of Appendix F, either X_1 and/or X_n is the outlier depending on the test statistic.

E.4.3.2. Example. Upon examining the ordered sample data: 82, 87, 92, 98, 103, 105, 106, 108, 113, and 151 mg/L, one suspects that the largest value (151 mg/L) of the data set could be an outlier. Because the mean value and standard deviation are sensitive to data distribution, a normality test is first performed on the data set. A normal probability plot shows that there is no reason to suspect that the data without suspected outliers are not normally distributed. Based on the mean value of 104.5 mg/L and standard deviation of 19.04 mg/L:

$$G = \frac{X_n - \bar{X}}{SD} = \frac{151 - 104.5}{19.04} = 2.44$$

Because G = 2.44, which is greater than 2.176 from Table F-1 of Appendix F, there is evidence that the largest value, 151, is an outlier at a 0.05 significance level and should be further investigated.

E.5. Homogeneity Test.

E.5.1. Procedure. To test the homogeneity of bulk PE sample material in a container, follow these steps:

1. Collect five samples randomly from each of the top, middle, and bottom sections of the container to yield a total of 15 samples. The globe null hypothesis is as follows:

$$H_O: \; :_1 = :_2 = :_3 = \ldots = :_{15}$$

where :'s are the population means that can be represented by the samples. The alternative hypothesis is:

H_A: The :'s are not all equal.

2. Carry out an ANOVA F test at " = 0.05 to test the globe null hypothesis following these steps:

- Calculate the sum and mean using following formulas.

$$sum = \sum_{j=1}^{n_i} y_{ij}$$

$$mean = \bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

where: y_{ij} = sample j in section i
n_i = number of samples in section i
sum = total amount of analyte in section i
mean = mean amount of analyte in section i

- Calculate the ANOVA quantities of within, between, and total Sum of Squares, *df*'s, and Mean Squares using following formulas.

$$Sum\ of\ Squares\ (within) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$Sum\ of\ Squares\ (between) = \sum_{i=1}^{m} n_i (\bar{y}_i - \bar{y})^2$$

$$Sum\ of\ Squares\ (total) = Sum\ of\ Squares\ (within) + Sum\ of\ Squares\ (between)$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

$$\bar{y} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{m} n_i}$$

$$df\ (within) = (\sum_{i=1}^{m} n_i) - m$$

$$df\ (between) = m - 1$$

$$df\ (total) = df\ (within) + df\ (between) = (\sum_{i=1}^{m} n_i) - 1$$

$$Mean\ Square\ (within) = \frac{Sum\ of\ Squares\ (within)}{df\ (within)}$$

$$Mean\ Square\ (between) = \frac{Sum\ of\ Squares\ (between)}{df\ (between)}$$

where: m = number of sections
df = degrees of freedom
Sum of Squares (within) = within-section variability
Sum of Squares (between) = between-section variability
Sum of Squares (total) = total variability
Mean Square = Sum of Squares divided by df

- Calculate the test statistic of F test as follows:

$$F = \frac{Mean\ Square\ (bewteen)}{Mean\ Square\ (within)}$$

- Obtain critical values from an F distribution (Table F-2 of Appendix F) with "Numerator df" and "Denominator df" equal to df (between) and df (within), respectively.

3. Interpret results: If F is less than the critical value at a significance level of 0.05, the material is considered homogeneous. If F is greater than the critical value but the total standard deviation of samples, i.e., the square root of Sum of Squares (total), is less than 0.3F, where F is the target standard deviation of the proficiency test, the material may still be regarded as sufficiently homogeneous. Otherwise, the material must be reprocessed and rechecked for homogeneity, or an alternative material selected. The only other approach would be to relax the target standard deviation for that particular material, i.e., to take account of the variance that the material will contribute to the results of individual participant laboratories. However, such a practice would destroy the utility of the proficiency test and the confidence of

participating laboratories and is not encouraged except for minor deviations from sufficient homogeneity.

E.5.2. Example. Five samples are randomly collected from each of the top, middle, and bottom sections of bulk PE sample material. Each sample is independently, randomly analyzed in duplicate and the mean values of the duplicate results are tabulated in Table E-2.

Table E-2. Example of Homogeneity Test

Section	Top	Middle	Bottom
Sample Data	85.5	84.0	79.0
	84.5	87.0	89.5
	83.5	82.0	82.5
	79.5	89.5	87.0
	83.0	86.5	85.0
n_i	5	5	5
sum	416	429	423
mean	83.2	85.8	84.6

Calculate the ANOVA quantities for analytes in the three sections and present them in an ANOVA table. The values in Table E-3 can be calculated with a calculator or statistical software.

Table E-3. ANOVA Table for Homogeneity Test

Source	Sum of Squares	*df*	Mean Square
Between sections	16.95	2	8.48
Within sections	119.80	12	9.98
Total	136.75	14	

Calculate the F test statistic as follows:

$$F = \frac{Mean\ Square\ (between)}{Mean\ Square\ (within)} = \frac{8.48}{9.98} = 0.850$$

Because F = 0.850 is less than the critical value 3.89 at a significance level of 0.05 with *df* (between) = 2 and *df* (within) = 12, H_O is not rejected. There is insufficient evidence to conclude any heterogeneity among different sections of PE sample material with respect to the mean concentration. The observed differences in mean concentrations can readily be attributed to chance variation.

E.6. Stability Test.

E.6.1. Test Types. Two approaches will be used to test the stability of PE samples. For PE samples of low stability and short holding times, The following test procedure applies:

1. Conduct duplicate analysis of at least five samples, randomly selected from the production run at the beginning and end of the proficiency test period.
2. Check the equality of variances of two population means with an appropriate statistical test, such as F test, Bartlett's test, Levene's test, etc.
3. If the variances of the two populations are approximately equal, compare the results at the end of the test period with the results at the beginning of the test period using a conventional two-sample *t* test. The mean values shall not be statistically different at " = 0.05 level.
4. If the two populations have unequal variances, compare the means with other tests such as Satterthwaite's two-sample *t* test.

For PE samples of high stability and longer holding time, use a trend test. See Section E.6.5 for procedures and description. The remainder of this section describes F Test, two-sample *t* test, and Satterthwaite's two-sample *t* test.

E.6.2. F Test.

E.6.2.1. Purpose and Underlying Assumptions. Use an F test to determine if the underlying variances of two populations are equal prior to a two-sample *t* test for equality of means. The assumptions underlying an F test are that the two samples are independent, random samples of normal distributions. Confirm these assumptions in the following ways:

- To determine if the samples are independent and random, review the sampling procedures.
- To determine normality, consider sample size. If both sample sizes are large, assume normality without further verifications. For small sample sizes, test the normality of each sample following the procedures in Section E.3.
- Nevertheless, a two-sample *t* test is robust to deviation from the normality of samples and equality of variances. In addition, because the stability test is usually performed by one laboratory, the variances of the two populations are usually equal.

E.6.2.2. Procedure. To perform an F test, following these steps:

1. Calculate the sample variances, SD_1^2 and SD_2^2, and the test statistic, F ratio, where SD_A^2 is the larger of the two variances.

$$F = \frac{SD_A^2}{SD_B^2}$$

2. Compare F with the critical $F_{1-\alpha/2}$ value at numerator $df = (n_A - 1)$ and denominator $df = (n_B - 1)$ in Table F-2 of Appendix F.
3. Interpret results: If $F < F_{1-\alpha/2}$, there is no evidence that the two variances of the two populations are different.

E.6.2.3. Example. Consider the data from a stability test below. Assume that each sample is independently and randomly collected and analyzed in duplicate. The mean values of the duplicate results are listed in Table E-4.

Table E-4. Example of Stability Test

Sample	Control (X_1) (Time t_0)	Test (X_2) (Time t_x)
	84.0	85.5
	87.0	84.5
	82.0	83.5
	89.5	79.5
	86.5	83.0
n_i	5	5
mean	85.8	83.2
SD_i	2.89	2.28
SD_i^2	8.35	5.20

where:
n_i = sample size (i.e., the number of X_i's)
mean = mean value of samples
SD_i = standard deviation of samples
SD_i^2 = variance of samples

$$F = \frac{SD_A^2}{SD_B^2} = \frac{SD_1^2}{SD_2^2} = \frac{8.35}{5.20} = 1.61$$

From Table F-2 of Appendix F, one determines the critical value, $F_{0.975}$, is 9.60 for numerator *df* = 4 and denominator *df* = 4 at a 0.05 significance level. Since F is less than $F_{0.975}$, there is no evidence that these two variances are different at 95% confidence level.

E.6.3. Student's Two-Sample *t* Test.

E.6.3.1. Purpose and Underlying Assumptions. Student's two-sample *t* test is used to compare the means of two populations when the variances of the two populations are equal. The basic assumptions required for the two-sample *t* test are independent and random sampling and normally distributed data. The two-sample *t* test is robust to violations of the assumptions of normality and equality of variances. However, if the assumptions of normality and equality of variances have been tested and rejected, use nonparametric methods such as Wilcoxon Rank Sum Test. Because sample means and standard deviations are used in the test, a two-sample *t* test is not robust to outliers.

E.6.3.2. Procedure. Follow these steps to complete a two-sample *t* test:

1. Use at least five samples, randomly selected from the production run at the beginning and end of a proficiency testing study. The null hypothesis is: H_O: $:_1 ' :_2$, where $:_1$ and $:_2$ are the mean analyte concentrations at time 1 (Time t_0) and 2 (Time t_x). The alternative hypothesis is: H_A: $:_1 \ldots :_2$. A conventional two-sample *t* test of the null hypothesis is normally carried out with $" = 0.05$.
2. Analyze each sample in duplicate randomly and independently.
3. Calculate the means and standard deviations of the duplicate results. See the following example in Table E-5.

Table E-5. Example of Student's Two-Sample *t* Test

Sample	Control (X_1) (Time t_0)	Test (X_2) (Time t_x)
	84.0	85.5
	87.0	84.5
	82.0	83.5
	89.5	79.5
	86.5	83.0
n_i	5	5
mean	85.8	83.2
SD_i	2.89	2.28

4. Be sure the variances of the two populations are equal. Because the stability testing is usually performed by one laboratory, it would be expected that sample data are of equal or nearly equal n_i and variances. The equality of variances can be checked with an F test as addressed in Section E.6.2. If the variances of the two samples are not equal, use Satterthwaite's *t* test (see Section E.6.4).
5. Compute the pooled standard deviation (SD_P) as follows. (SD_P below uses SD_i values from the example table.)

$$SD_P = \sqrt{\frac{(n_1 - 1)\, SD_1^2 + (n_2 - 1)\, SD_2^2}{n_1 + n_2 - 2}} = 2.60$$

where: n_i = sample size (i.e., the number of X_i's)
SD_i = standard deviation of samples

6. Compute the *t* test value and degrees of freedom using these formulas. (Computed *t* value uses SD_P calculated previously. Degrees of freedom is calculated using data from the example.)

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{SD_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 1.58$$

$$df = (n_1 + n_2 - 2) = 8$$

7. Use Table F-3 of Appendix F to find that the $t_{0.025}$ (i.e., two-tailed 5% critical value) with $df = 8$ is 2.306. Because t Ý $t_{0.025}$, there is not enough evidence to reject the null hypothesis, i.e., the data do not provide sufficient evidence to claim that the PE samples are unstable.

E.6.4. Satterthwaite's *t* Test.

E.6.4.1. Purpose and Underlying Assumptions. Use Satterthwaite's *t* test to compare the means of two populations when the variances of the two populations are not equal. The test is demonstrated below with the same example as the two-sample *t* test (assuming unequal variances).

1. Use at least five samples, randomly selected from the production run at the beginning and end of the proficiency testing study. The same null hypothesis applies: H_O: $\mu_1 = \mu_2$, where μ_1 and μ_2 are the mean analyte concentrations at time 1 (Time t_0) and 2 (Time t_x). The alternative hypothesis is: H_A: $\mu_1 \neq \mu_2$. Satterthwaite's *t* test with $\alpha = 0.05$ is carried out to test the null hypothesis.
2. Analyze each sample in duplicate independently and randomly.
3. Calculate the mean values of the duplicate results.

See the following example of Satterthwaite's *t* test in Table E-6.

Table E-6. Example of Satterthwaite's *t* Test

Sample	Control (X_1) (Time t_0)	Test (X_2) (Time t_x)
	84.0	85.5
	87.0	84.5
	82.0	83.5
	89.5	79.5
	86.5	83.0
n_i	5	5
mean	85.8	83.2
SD_i	2.89	2.28
SE_i	1.29	1.02

where: n_i = sample size (i.e., the number of X_i's)
mean = mean value of samples
SD_i = standard deviation of samples
SE_i = standard error of mean

The standard error of the mean, SE_i, is defined as:

$$SE_i = \frac{SD_i}{\sqrt{n_i}}$$

4. Calculate the standard error of the difference between two sample means using the formula below.

$$SE_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{SE_1^2 + SE_2^2} = \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}} = 1.65$$

5. Compute the test statistic for Satterthwaite's *t* test using the formula below. (Computed *t* value uses *SE* calculated previously. Degrees of freedom is calculated using data from the example.)

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{SE_{(\bar{X}_1 - \bar{X}_2)}} = \frac{(85.8 - 83.2)}{1.65} = 1.58$$

$$df = \frac{(SE_1^2 + SE_2^2)^2}{\frac{SE_1^4}{(n_1 - 1)} + \frac{SE_2^4}{(n_2 - 1)}} = 7.56$$

6. Determine the $t_{0.025}$ (i.e., two-tailed 5% critical value) of Student's *t* distribution with $df = 7.56$ by computer software or estimate it from Table F-3 of Appendix F, with *df* given by the smaller of $(n_1 - 1)$ and $(n_2 - 1)$, or by $(n_1 + n_2 - 2)$. The former estimate is somewhat conservative (i.e., the true confidence level is slightly greater than 95% when $t_{0.025}$ is used), and the later estimate is somewhat liberal; however, both usually have little effect on the final decision. If $df = (n_1 + n_2 - 2) = 8$, the critical value, $t_{0.025}$, is 2.306, which is greater than the observed *t* value of 1.58. Because the significance level of the data is greater than $\alpha = 0.05$, the data are judged compatible with H_O; H_O is not rejected, and the data do not provide sufficient evidence to claim that the PE samples are unstable.

E.6.5. Trend Test. Specific procedures, calculation method, and an example of statistical trend tests follows in this section.

E.6.5.1. Procedure. For PE samples of high stability and longer holding time such as real-world PE samples, use an alternative approach to check the stability of PE sample material. Follow these guidelines:

1. Analyze in duplicate at least five samples, randomly selected from different sections of the bulk material.
2. Perform the analysis in the beginning and at the end of the proficiency test period.
3. Compare the results at the end of the test period with the results from the beginning of the test period. The mean shall not be statistically different at " = 0.05 level, using a conventional *t* test as previously described.
4. Test the bulk material at a regular time interval over an extended time period to cover the anticipated shelf-life or until the supply of the PE sample material is exhausted. (Base the anticipated shelf-life on prior experience and/or evidence from technical literature.)

5. Control chart test data and use to determine and monitor the stability, according to ISO 7870, "Control Charts - General Guide and Introduction;" ISO 7966, "Acceptance Control Charts;" ISO 8258, "Shewhart Control Charts;" and ASTM, "Manual on Presentation of Data and Control Chart Analysis," 1990.

E.6.5.2. Statistical Trend Tests. These tests, such as the Mann-Kendall trend test, are very helpful in detection of any trends in measured concentrations over time and in prevention of potential problems with sample contamination or degradation. The basic Mann-Kendall trend test involves listing the sample data in a temporal order, and computing all differences that may be formed between each datum and its earlier data across a triangular table. If there is an underlying upward or downward trend, then these differences will tend to be sufficiently large positive or negative values to suggest the presence of an upward or downward trend, respectively. Zero differences are not included in the test statistic and therefore should be avoided, if possible, by recording data to sufficient accuracy. For censored data of nondetects, a value of half of MDL value shall be assigned to data reported as below the MDL. The basic Mann-Kendall trend test could be applied to small sample sizes (i.e., fewer than ten). For larger sample sizes, a normal approximation to Mann-Kendall test can be used and is described in a number of textbooks. The procedure and an example of the basic Mann-Kendall trend test are presented below.

E.6.5.3. Mann-Kendall Calculation. The Mann-Kendall statistic, S, is calculated as follows:

1. Calculate the total number of positive signs minus the total number of negative signs across in the triangular table.
2. Determine probability "p" using Table F-4 of Appendix F, sample size n, and absolute value of Mann-Kendall statistic S.

3. Interpret results: A positive S indicates a potential upward trend and a negative S indicates a potential downward trend. For an upward trend, reject H_O (i.e., no trend) if S > 0 and p < ". For a downward trend, reject H_O if S < 0 and p < ".

E.6.5.4. Example. Consider the following five stability test data points listed in a chronological order: 80, 64, 88, 48, and 50 mg/L. The null hypothesis, H_O: No trend, versus the alternative hypothesis H_A: Either an upward or downward trend at " = 0.05 significance level is used for a trend test of sample stability. After examining the data listed below in Table E-7, the suspected trend is obviously a downward trend. The triangular table below is constructed to list all possible differences. The sum of signs of the differences across the rows is shown in the last two columns.

Table E-7. Example of Trend Test

Time	1	2	3	4	5	No. of	No. of
Data	80	64	88	48	50	"+" Signs	" ! " Signs
80		!	+	!	!	1	3
64			+	!	!	1	2
88				!	!	0	2
48					+	1	0
						3	7

The Mann-Kendall statistic S = (number of "+" signs) ! (number of " ! " signs) = 3 ! 7 = !4. Based on Table F-4 of Appendix F, the probability, p, is 0.242 with *n* = 5, and *S* = 4. Because S = !4 < 0 and p = 0.242 Û " (0.05), the null hypothesis is not rejected and there is not sufficient evidence to conclude that there is a downward trend in concentration or instability of the PE sample at a 0.05 significance level.

E.7. Reproducibility.

E.7.1. Procedure. A within-batch reproducibility test is similar to a homogeneity test. It consists of the following steps:

1. Run independent tests of multiple PE samples randomly selected from a production batch. Test each sample in duplicate or triplicate, yielding a minimum of 15 tests.
2. Use a traditional ANOVA F test to test the significance of within-batch component of variances as for a homogeneity test.
3. Interpret results: If the F test is not significant at " = 0.05 level, the PE samples may be considered equivalent and the production process reproducible. If the F test is significant, conduct an additional test to assure that the significance represents a difference that truly could affect the evaluation of proficiency testing results. To do this, compare the size of

within-batch component of variances with the acceptance limits of PE samples. If the within-batch component is less than 10% of the acceptance limits (i.e., 3F), the within-batch PE samples are considered sufficiently reproducible.

E.8. Reference Value.

E.8.1. Procedure. A number of methods are available to PE Sample Suppliers to establish the reference value for the analyte. They are similar to the procedures required to assign a concentration value to a reference material and are described below:

1. Set the reference value at the mean value of interlaboratory study if the mean value is within the method-specified control limits on bias. The value should be compatible with the prepared value and/or the mean of referee laboratories with a conventional *t* test at " = 0.05 significance level, depending on the specific type of analytes and matrices. (See Section 5.4 of Chapter 5 for procedures to determine reference values of PE samples.)
2. Determine if the underlying population distribution is single modal with a normal distribution prior to computing mean value. Use graphic presentations such as a histogram plot, frequency plot, stem-and-leaf plot, ranked data plot, quantile plot, normal probability plot, etc. to determine this.
3. Perform an outlier test to determine if an extreme value is a statistical outlier and if the statistical outlier should be excluded or modified prior to calculation of mean value. (A mean value is sensitive to extreme values and nondetects.)
4. Set concentrations for nondetects at half of the sample-specific detection limits if the limits meet the study requirements.
5. Perform an outlier test to decide if the estimated value for a nondetect should be included for computation of sample mean.

E.8.3. Median. Sample median (i.e., 50% point) has recently become very popular in replacing sample mean when robust statistics is used. Sample median is not sensitive to extreme values and can easily be used to handle censored data, i.e., nondetect. The USACE PE Sample Program is not currently using robust statistics, but may consider it for specific PE samples on a case-by-case basis. The discussion of robust statistics is beyond the scope of this Appendix.

E.9. Prediction Intervals.

E.9.1. Definition. A sample mean is an estimate of the unknown population mean, :, but differs from it because of sampling fluctuations. However, it is possible to construct a statistical interval known as a confidence interval to contain the population mean with a specific probability which is known as the associated confidence level. Thus a 95% confidence interval on the population mean is an interval which contains : with a probability of 0.95, or, over a large number of samples, the 95% confidence interval will contain the unknown population mean 95%

of the time. Most textbooks on statistics devote extensive space to confidence intervals on population parameters. A prediction interval estimates the range of variation of the observations in a future sample and is used to obtain limits to contain all of a small number of future data based on previous *n* data and a specified probability. In proficiency testing, it is the interval within which the next laboratory performance is expected to be located based on the results of *n* prior laboratory proficiency tests. For the USACE PE Program, the prediction intervals that will contain the next proficiency test result at 95% and 99% confidence intervals are used to establish the acceptance limits for proficiency testing.

E.9.2. Calculation. Assume that *n* independent and random data are available. The two-sided prediction interval can be constructed from the sample mean and standard deviation (*SD*) as follows.

$$\bar{X} \pm k(n,\ 1-\alpha) \times SD$$

where: $k(n,\ 1-\alpha)$ in Table F-5 of Appendix F is a factor for determining two-sided $100\times(1-\alpha)\%$ prediction intervals for a single future observation given a previous sample of size *n*.

For values not tabulated in Table F-5, a conservative approximation for $k(n,\ 1-\alpha)$ is:

$$k(n,\ 1-\alpha) = \sqrt{1 + \frac{1}{n}} \times t(n-1,\ 1-\alpha/2)$$

where: n = size of previous samples
$t(n-1,\ 1-\alpha/2) = [100\times(1-\alpha/2)]$'th percentile of the Students' *t* distribution with $(n-1)$ degree of freedom

This approximation works satisfactorily for most practical purposes, except for combinations of small *n* and large α. If the *n* previous data and the single future datum are all randomly selected from the same normal distribution, one can state with $(1-\alpha)\%$ confidence that the single future datum will be within the calculated prediction interval. A prediction interval is sensitive to the assumption of normality of the data distribution. Other procedures for constructing prediction intervals that make no assumptions about the distribution type are also available; however, they are beyond the scope of this Appendix.

E.9.3. Examples. Consider the following data obtained on a normally distributed parameter based on a random sample from an interlaboratory study: 50.9, 45.8, 49.1, 46.0, and 50.4 ：g/L. The sample mean is 48.44 ：g/L, and the standard deviation is 2.41 ：g/L. The 95% and 99% prediction intervals are:

95%: $48.44 \pm (3.041)(2.41) = 48.44 \pm 7.33$

99%: $48.44 \pm (5.043)(2.41) = 48.44 \pm 12.15$

Thus the next datum from the sample population should fall in an interval of 41.11 ：g/L to 55.77 ：g/L with a probability of 0.95, and in the interval of 36.29 ：g/L to 60.59 ：g/L with a probability of 0.99.